Title

Multivariate adaptive regression splines for estimating riverine constituent concentrations

Journal

Hydrological Processes, 34(5)

ISSN

0885-6087

Authors

Huang, H
Ji, X
Xia, F
et al.

Publication Date

2020-02-28

DOI

10.1002/hyp.13669

Peer reviewed

Received: 11 June 2019 | Accepted: 1 November 2019
DOI: 10.1002/hyp.13669

RESEARCH ARTICLE

WILEY

Multivariate adaptive regression splines for estimating riverine constituent concentrations

Hong Huang[1,2] | Xiaoliang Ji[2,3] | Fang Xia[2,3] | Shuhui Huang[1,2] | Xu Shang[2,3] | Han Chen[2] | Minghua Zhang[2,3] | Randy A. Dahlgren[2,4] | Kun Mei[1,2]

[1]Health Assessment Center, Wenzhou Medical University, Wenzhou, China

[2]Zhejiang Provincial Key Laboratory of Watershed Science and Health, School of Public Health and Management, Wenzhou Medical University, Wenzhou, China

[3]Southern Zhejiang Water Research Institute, Wenzhou, China

[4]Department of Land, Air, and Water Resources, University of California, Davis, California

Correspondence
Kun Mei, Health Assessment Center, Wenzhou Medical University, Wenzhou 325035, Zhejiang Province, China.
Email: meikun@iwaterlab.com

Funding information
National Natural Science Foundation of China, Grant/Award Numbers: 41601554, 41807495

Abstract

Regression-based methods are commonly used for riverine constituent concentration/flux estimation, which is essential for guiding water quality protection practices and environmental decision making. This paper developed a multivariate adaptive regression splines model for estimating riverine constituent concentrations (MARS-EC). The process, interpretability and flexibility of the MARS-EC modelling approach, was demonstrated for total nitrogen in the Patuxent River, a major river input to Chesapeake Bay. Model accuracy and uncertainty of the MARS-EC approach was further analysed using nitrate plus nitrite datasets from eight tributary rivers to Chesapeake Bay. Results showed that the MARS-EC approach integrated the advantages of both parametric and nonparametric regression methods, and model accuracy was demonstrated to be superior to the traditionally used ESTIMATOR model. MARS-EC is flexible and allows consideration of auxiliary variables; the variables and interactions can be selected automatically. MARS-EC does not constrain concentration-predictor curves to be constant but rather is able to identify shifts in these curves from mathematical expressions and visual graphics. The MARS-EC approach provides an effective and complementary tool along with existing approaches for estimating riverine constituent concentrations.

KEYWORDS

concentration-discharge curve, concentration-season curve, pollutant flux, uncertainty analysis, water quality, watershed management

1 | INTRODUCTION

Water quality degradation, particularly to drinking water sources and aquatic habitats, is a major global concern (Calamari, Nauen, & Naeve, 1987; Dumont, Williams, Keller, VoÃ, & Tattari, 2012; Huang, Chen, Zhang, Zeng, & Dahlgren, 2014; León, Soulis, Kouwen, & Farquhar, 2001; Ouyang, Cai, Huang, & Hao, 2016). Riverine constituent concentration and load estimation is a primary data requirement for guiding basic water quality protection practices, water quality risk assessment, and watershed management and remediation (Huang et al., 2017). Currently, the common methods for riverine concentration and load estimation can be divided into two categories of (a) mechanistic model-based methods, such as the soil and water assessment tool and hydrological simulation program—Fortran (Saleh & Du, 2004), and (b) regression-based methods, such as the multiple log linear regression model (ESTIMATOR; Cohn, Delong, Gilroy, Gilroy, & Wells, 1989), and the weighted regressions on time, discharge, and season (WRTDS; Hirsch et al., 2010). Mechanistic models generally require considerable time and effort and may be impractical because they have large data requirements, complex structure, and require large calibration parameter sets that are difficult to estimate at the watershed scale (Chen, Dahlgren, & Lu, 2013). In contrast,

Hong Huang and Xiaoliang Ji contributed equally to this work.

regression-based methods require discrete paired samples of concentration and discharge for parameter estimation and continuous records of discharge to estimate concentrations for a desired period (Grizzetti, Bouraoui, de Marsily, & Bidoglio, 2005; Huang, Zhang, & Lu, 2014).

From a statistical perspective, current regression-based methods can be further divided into parametric and nonparametric methods, whose prototypical representations are ESTIMATOR and WRTDS, respectively. A common assumption of regression-based methods is that riverine concentrations are a function of several factors, such as time, discharge, and season. ESTIMATOR is a parametric linear multivariate regression model that predicts constituent concentration or load by developing a linear relationship between predictor variables, such as ln discharge, ln discharge^2, time, time^2, and season (Cohn, Caulder, Gilroy, Zynjuk, & Summers, 1992). However, ESTIMATOR assumes that the concentration-season and concentration-discharge curves are constant through time. To overcome this shortcoming, Yochum (2000) suggested using estimation windows (e.g., 9 years), yet the window length selection process lacks a rigorous theoretical basis and is therefore subjective (Appling, Leon, & McDowell, 2016). By comparison, WRTDS does not require any underlying mathematical form and makes a three-dimensional matrix of locally weighted regressions in time–discharge–season (Hirsch & Cicco, 2015; Shipley & Hunt, 1996). The most significant advantage of WRTDS may be its capacity to allow concentration-season and concentration-discharge curves to change through time (Hirsch, Archfield, & Cicco, 2015; Hirsch & Cicco, 2015). If the advantages of parametric and nonparametric regression-based methods could be integrated in a new approach, it would provide valuable and practical significance to a wide range of water quality studies.

Multivariate adaptive regression splines (MARS) is a method for flexible modelling of high dimensional data (Friedman, 1991). MARS does not impose any specific relationship type between the response variable and predictor variables but takes the form of an expansion in product spline functions, where the number of spline functions and interactions are automatically determined by the data (Friedman & Roosen, 1995). Because of these beneficial features, MARS has strong pattern detection ability, which has been widely used in fields such as ecology (Leathwick, Elith, & Hastie, 2006), medicine (Koba & Bączek, 2013), and economics (Lorca & Juez, 2011). In recent years, MARS was successfully applied in hydrology and water resources, such as for drought forecasting (Deo, Kisi, & Singh, 2017) and evaporation modelling (Kisi, 2015). MARS can be considered a semiparametric method that can often fill the gap between parametric and nonparametric methods and therefore has potential as a new approach for estimating riverine constituent concentrations and fluxes.

The U.S. Geological Survey (USGS) monitors water quality in Chesapeake Bay watershed at nine long-term River Input Monitoring stations located on the Susquehanna, Potomac, James, Rappahannock, Appomattox, Pamunkey, Mattaponi, Patuxent, and Choptank Rivers. Since 1985, the USGS has collected a minimum of 20 samples per year at each of the nine River Input Monitoring stations, and the samples are collected across the full range of the hydrologic conditions, including 12 monthly samples and eight targeted storm flow samples (i.e., periods of elevated discharge; Moyer, Hirsch, & Hyer, 2012). These datasets were selected for their several beneficial attributes, such as being long term with consistent analytical methods and QC/QA protocols, nonmonotone data distribution, and sufficiently high sampling frequencies. Furthermore, this dataset was open access and has been used in previous studies (Brakebill, Scott, & Schwarz, 2010; Moyer et al., 2012; Zhang, Brady, & Boynton, 2015; Zhang, Hirsch, & Ball, 2016). Description of the rivers and water quality parameters is provided in the Supplementary Information A. The aim of this paper is (a) to introduce a new semiparametric method called MARS-EC for estimating river constituent concentrations from river discharge records, (b) to demonstrate the process, interpretability and flexibility of the MARS-EC modelling approach, and (c) to assess model accuracy and uncertainty. The performance of MARS-EC was evaluated using river water quality and discharge datasets from the USGS monitoring program in Chesapeake Bay.

2 | METHODOLOGIES

2.1 | A brief introduction to MARS

Objectively speaking, MARS is a nonparametric regression technique but can be used as an adaptive non-linear regression that uses piecewise functions to define relationships between a response variable and multiple predictors. The model form of MARS is

$$\hat{f}(x) = c_0 + \sum_{i=1}^{k} c_i B_i(x) + \varepsilon, \quad (1)$$

where c_0 is a constant representing the model intercept, $B_i(x)$ is a basic function (spline function), k is the number of basic functions, c_i is the constant coefficient of $B_i(x)$, and ε is the unexplained variation.

In MARS models, each basic function takes the form of a constant (i.e., the intercept) and a *hinge function*, as well as a product of two or more *hinge functions* to model interactions between variables if necessary. *Hinge functions* are a key component of MARS models and take the form of

$$\begin{gathered} \max(0, x-c) \\ \text{or} \\ \max(0, c-x), \end{gathered} \quad (2)$$

where c is a constant called a *knot* (i.e., a break point value), and max $(0, x - c)$ or max$(x - c, 0)$ refer to *hinge functions* where max$(0, x - c)$ is 0 if $x - c < 0$, else $x - c$.

MARS builds a model in two phases using a forward and backward pass. The forward pass usually builds an overfit model. In the backward pass, the generalized cross validation (GCV) criterion is used to find the overall best model from a sequence of fitted models, where a larger GCV value tends to produce a smaller model, and vice versa (Oduro, Metia, Duc, Hong, & Ha, 2015). The GCV is use to

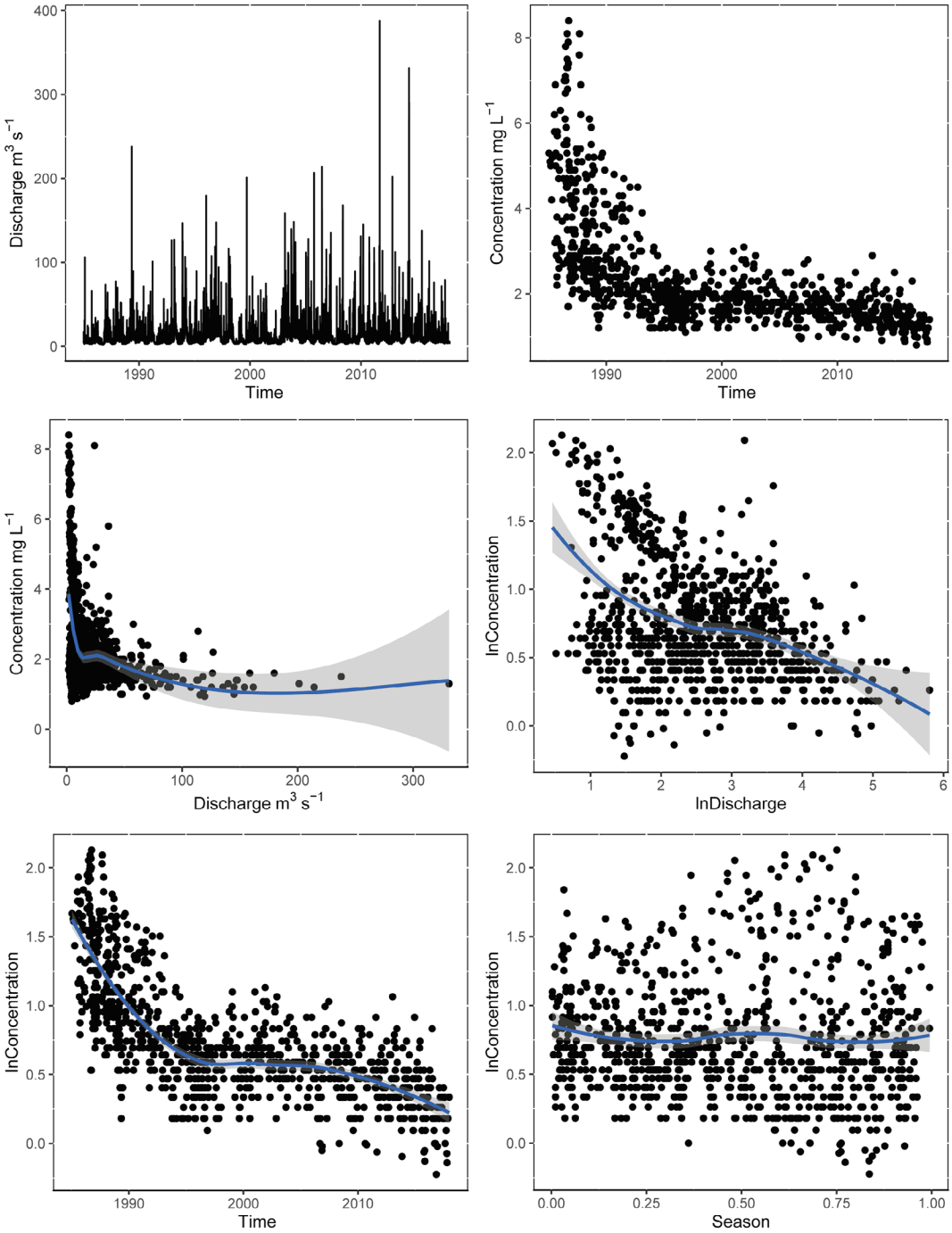

FIGURE 1 Concentration and predictor variables for TN in the Patuxent River (1985–2017). The smooth curve is the loess smooth line and the shadow is the standard error

achieve a balance between model fitting ability and model complexity (Friedman & Roosen, 1995):

$$GCV = RSS/\left(N(1-ENP/N)^2\right), \quad (3)$$

where *RSS* is the residual sum-of-squares measured on the training data, *ENP* is the effective number of parameters, and *N* is the number of observations.

The effective number of parameters is defined as

$$ENP = NMT + Penalty(NMT-1)/2, \quad (4)$$

where *NMT* is the number of MARS terms and *Penalty* is about 2 or 3.

Note that (Number of MARS Terms - 1)/2 is the number of hinge-function knots, so the formula penalizes the addition of knots, and therefore, the GCV formula adjusts (i.e., increases) the training *RSS* to

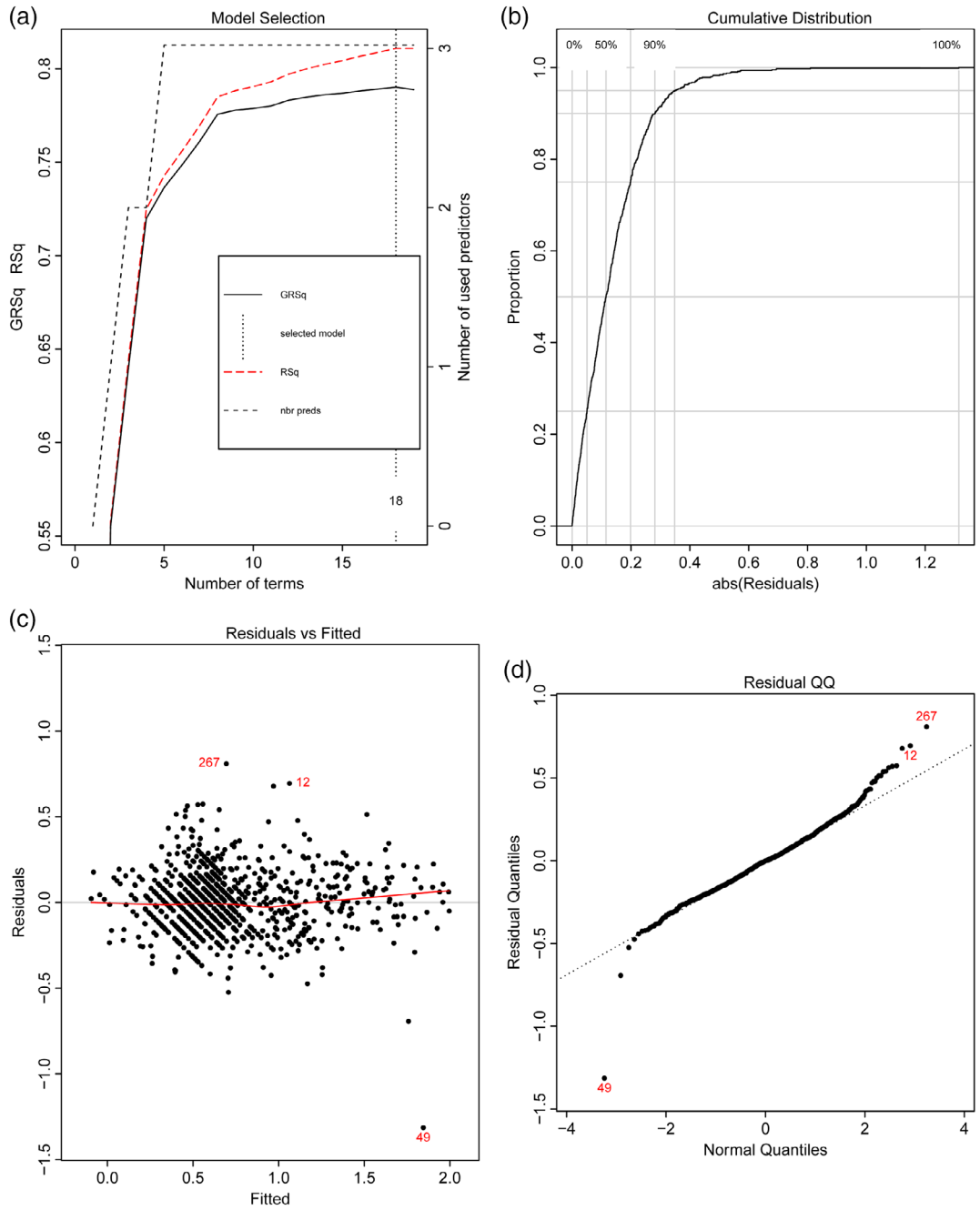

FIGURE 2 Model evaluation metrics of MARS-EC model for TN in Patuxent River (1985–2017). (a) Model selection graph, (b) residuals versus fitted graph, (c) cumulative distribution graph, and (d) residual QQ graph

take into account the flexibility of the model (Fazel Zerandi, Zarinbal, Ghanbari, et al., 2013).

2.2 | Mars-EC

In the MARS-EC approach, riverine constituent concentrations are assumed to be functions of environmental factors such as time, discharge, and season, as well as other environmental factors. The predictor variables in MARS-EC are adopted from but not limited to those used in ESTIMATOR and WRTDS (Cohn et al., 1992; Hirsch et al., 2010). However, MARS-EC assumes riverine constituent concentrations to be piecewise linear functions of predictors and allows concentration-predictors such as concentration-discharge curves to change over time by means of interaction items. The mathematical expression of MARS-EC for concentration estimation is

$$\begin{aligned}\ln Concentration(T) = c_0 + &\sum_{i=1}^{k} c_i B_i Time(T) + \\ &\sum_{i=1}^{k} c_i B_i \ln Discharge(T) + \\ &\sum_{i=1}^{k} c_i B_i Season(T) + \\ &\sum_{i=1}^{k} c_i B_i EnvironFactor_i(T) + \cdots \varepsilon,\end{aligned} \quad (5)$$

where ln*Concentration*(*T*) is the logarithmic value of measured daily constituent concentration, c_0 is a constant representing the model intercept, *k* is the number of basic functions, c_i is a constant coefficient for each basic function, $\sum_i^k c_i B_i Time(T)$ represents the relationship between concentration and time (in decimal years), $\sum_i^k c_i B_i(\ln Discharge)$ represents the relationship between concentration and discharge, $\sum_i^k c_i B_i(Season)$ represents the relationship between concentration and season (day in year, in decimal days), $\sum_i^k c_i B_i Environ\_Factor_i(T)$ represents the relationship between concentration and environmental factor *i*, and ε is the unexplained variation.

2.3 | Model example

MARS-EC was developed in R programming language for statistical computing using the "earth" package (Version 4.6.3; Milborrow, 2018). The earth R package builds regression models using the techniques in Friedman's papers "Multivariate Adaptive Regression Splines" and "Fast MARS" (Milborrow, 2019a). To demonstrate the step-by-step modelling approach using MARS-EC, an example for estimating total nitrogen (TN; nitrate + nitrite + ammonia + organic-N) is provided for a 33-year record from the Patuxent River near Bowie, Maryland, a major river input to Chesapeake Bay. The main R codes for this model example are provided in Supplementary Information B.

2.3.1 | Data processing

In this model example, TN concentration is regressed using three predictor variables of Trend, Discharge, and Season, as well as potential interaction effects. Discrete TN concentration (about semi-monthly) and daily discharge data were downloaded from USGS Data Services (http://waterservices.usgs.gov/). Data were retrieved using the "data Retrieval" packages (Version 2.7.5) in R (DeCicco, Hirsch, & Lorenz, 2019). A data frame containing these dependent and predictor variables was set up in accordance with format requirements for "earth" packages. Visual plots were used to assess data patterns before the subsequent model selection step, such as the need to make log transformations of concentration and discharge data before model development (Figure 1).

2.3.2 | Model selection and residual evaluation

In MARS, the final model was selected at the maximum GCV, and the generalization ability of the model was assessed by RSq (i.e., the coefficient of determination) and GRSq statistics.

$$GRSq = 1\text{-}GCV/GCV.null, \quad (6)$$

where GCV.null is the GCV of an intercept-only model.

Upon completion, the "earth" packages produce graphs to describe model selection and evaluate performance. Commonly used graphs include "Model Selection," "Residuals vs Fitted," "Cumulative Distribution," and "Residual QQ". In our example, the best model had 18 terms and used all three predictors (Equation 7 and Figure 2a); GRSq and RSq were ~0.80 (Figure 2a). The "Cumulative Distribution" graph indicates that the distribution starts at 0 and shoots up quickly to 90% at 0.3 (Figure 2b). The "Residuals vs Fitted" and "Residual QQ" graphs highlight the cases (default is 3) having the largest residuals (Figure 2c,d). Although cases having large residuals could be excluded when building the model, they might reveal important data considerations that could warrant changes to the model (Milborrow, 2019a). However, dealing with potential outliers requires careful consideration and knowledge of the system, which were beyond the scope of this example. Additionally, "earth" packages provide "Abs residuals vs fitted," "Sqrt abs residuals vs fitted," "Abs residuals vs log fitted," "Cube root of the squared residuals vs log fitted," and "Log abs residuals vs log fitted" for user analysis (Milborrow, 2019b). MARS-EC also calculates the prediction and confidence intervals that are useful for uncertainty assessment (Buckner, Choi, & Gibson, 2006; Khosravi, Mazloumi, Nahavandi, Creighton, & Lint, 2011). In this example, the mean, smallest, and largest values of the 95% prediction interval were 0.72, 0.62, and 0.89, respectively, with 92% of the values falling into the 90% prediction intervals (Figure 3).

$$\begin{aligned}\ln Concentration = {} & 0.406 \\ & + 0.222\max(0,\ 1993.781 - \text{Trend}) \\ & + 1.588\max(0,\ 3.582 - \ln\text{Discharge}) \\ & - 0.187\max(0,\ \ln\text{Discharge} - 3.582) \\ & + 0.236\max(0,\ 0.838 - \text{Season}) \\ & + 1.196\max(0,\ \text{Season} - 0.838) \\ & - 0.073\max(0,\ \text{Trend} - 1989.800)\max(0,\ 3.582 - \ln\text{Discharge}) \\ & - 0.106\max(0,\ 1993.781 - \text{Trend})\max(0,\ \ln\text{Discharge} - 1.287) \\ & + 0.091\max(0,\ 1993.781 - \text{Trend})\max(0,\ \ln\text{Discharge} - 2.168) \\ & - 0.067\max(0,\ 2011.931 - \text{Trend})\max(0,\ 3.582 - \ln\text{Discharge}) \\ & + 0.030\max(0,\ \text{Trend} - 2011.931)\max(0,\ 3.582 - \ln\text{Discharge}) \\ & - 0.122\max(0,\ 1993.781 - \text{Trend})\max(0,\ \text{Season} - 0.595) \\ & - 0.109\max(0,\ 1993.781 - \text{Trend})\max(0,\ 0.595 - \text{Season}) \\ & + 0.325\max(0, 3.582 - \ln\text{Discharge})\max(0, \text{Season} - 0.688) \\ & + 0.172\max(0, 3.582 - \ln\text{Discharge})\max(0, 0.688 - \text{Season}).\end{aligned} \quad (7)$$

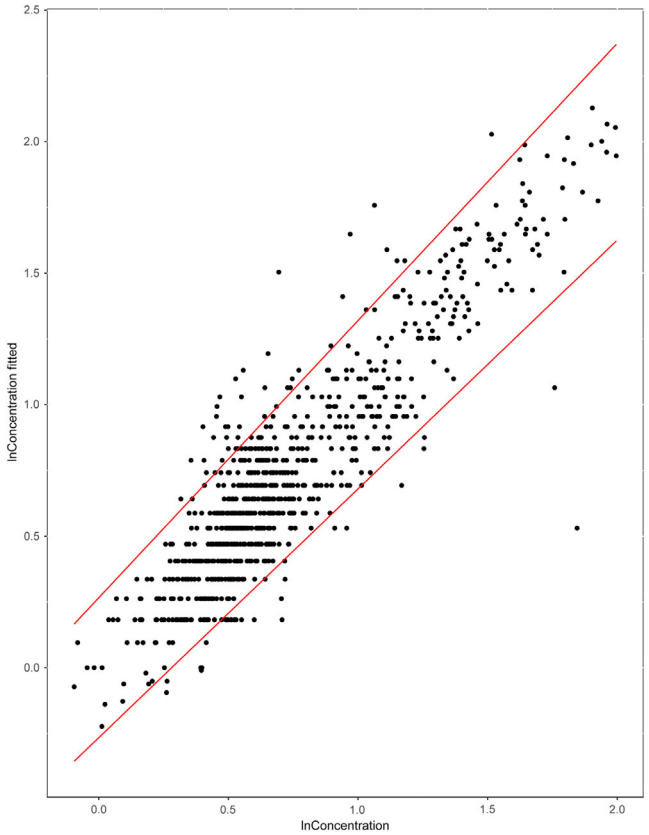

FIGURE 3 Prediction and 90% CI (red lines) of MARS-EC model for TN in Patuxent River (1985–2017)

2.3.3 | Results visualization

MARS provides mathematical expressions and graphs for visualization of results. For the TN concentration model in the Patuxent River, the mathematical expression with calibrated coefficients is shown in Equation (7). For the main effect, MARS-EC generates a separate graph to show the concentration-predictor relationship by holding all other variables at their median values. For interaction effects, MARS-EC plots changing concentrations for two variables while holding all other variables at their median values. In our example, the main effect and interaction effects are showed in Figure 4. TN concentration gradually decreased with time (Figure 4a) and change points were detected at 1989.800, 1993.781, and 2011.931 (Equation 7). For the concentration-discharge curve (Figure 4b), lnConcentration decreased with increasing lnDischarge, especially at lnDischarge greater than 3.583 (Equation 7). For the concentration-season curve (Figure 4a), lnConcentration decreased with the progression of seasons (i.e., day in year) from January to October and then showed an increase (after day in year >0.838; Equation 7). The main effects quantify each concentration-predictor curve on the whole, whereas the interaction effects identify changes in each concentration-predictor curve caused by other predictors. The interaction effect is essential to concentration estimation since the concentration-discharge and concentration-season curves can change over time (Hirsch et al., 2010; Moyer et al., 2012). Using the calibrated model and daily discharge, decimal date (trend pattern) and decimal day in year (season pattern), it is easy to calculate trends in concentrations for any time scale/parameter (Figure 5).

3 | RESULTS AND DISCUSSION

3.1 | Model interpretability

Model interpretability refers to the capability of the model to express the behaviour of the system in an understandable way (Casillas, Cordón, Herrera, & Magdalena, 2003). For riverine constituent concentration estimation, model interpretability mainly depends on the capability of the model to capture long-term trends and seasonal patterns, as well as the important influence of discharge. The trend pattern of water quality is very important in environmental water management and remediation (Chang, 2008; Huang et al., 2017). MARS-EC demonstrated a strong capacity for trend and change point analysis for water quality constituents. In MARS-EC, the trend pattern was calculated using the calibrated model and the same predictors, but all variables except trend item were held at their median values. Therefore, the trend patterns derived from MARS-EC could be considered as discharge-season-adjusted concentrations. This is a different approach from previous methods. For instance, in WRTDS, the flow-normalization uses the actual historical discharge values for a given day, with each historical value being assigned an equal probability of occurrence in any given year (Hirsch et al., 2010). In our example, trend patterns for yearly concentrations (Figure 6) showed three change points in 1989, 1993, and 2011. Therefore, we can say that the trend patterns for annual TN concentrations in the Patuxent River decreased from 4.26 mg L^{-1} in 1985 to 3.72 mg L^{-1} in 1989 (slope = −0.14 mg L^{-1} $year^{-1}$), more rapidly to 2.08 mg L^{-1} in 1993 (slope = −0.42 mg L^{-1} $year^{-1}$), more slowly to 1.64 mg L^{-1} in 2011 (slope = −0.02 mg L^{-1} y^{-1}), and finally to 1.09 mg L^{-1} in 2017 (slope = −0.09 mg L^{-1} $year^{-1}$). The MARS-EC approach determines coefficients for explanatory variables, the non-linear and nonmonotonic influences of predictor variables and their interaction effects can be quantitatively estimated and visually displayed. From this point of view, the MARS-EC approach provides an effective and useful tool for trend change detection (i.e., change point identification), which is an important attribute for devising water protection plans, as well as for assessing the effectiveness of ongoing watershed management activities (Chang, 2008; Huang et al., 2017; Renwick, Vanni, Zhang, & Patton, 2008).

A key goal in riverine constituent concentration and load modelling is to accurately capture the concentration-discharge and concentration-season relationships (Aulenbach et al., 2016; Hirsch, 2014). Another important capacity of MARS-EC is its ability to identify shifts in concentration-discharge and concentration-season curves. In accordance with the interaction plot between Trend and lnDischarge (Figure 4a), as well as the change points in Equation (7), the lnConcentration–lnDischarge relation curves showed distinct

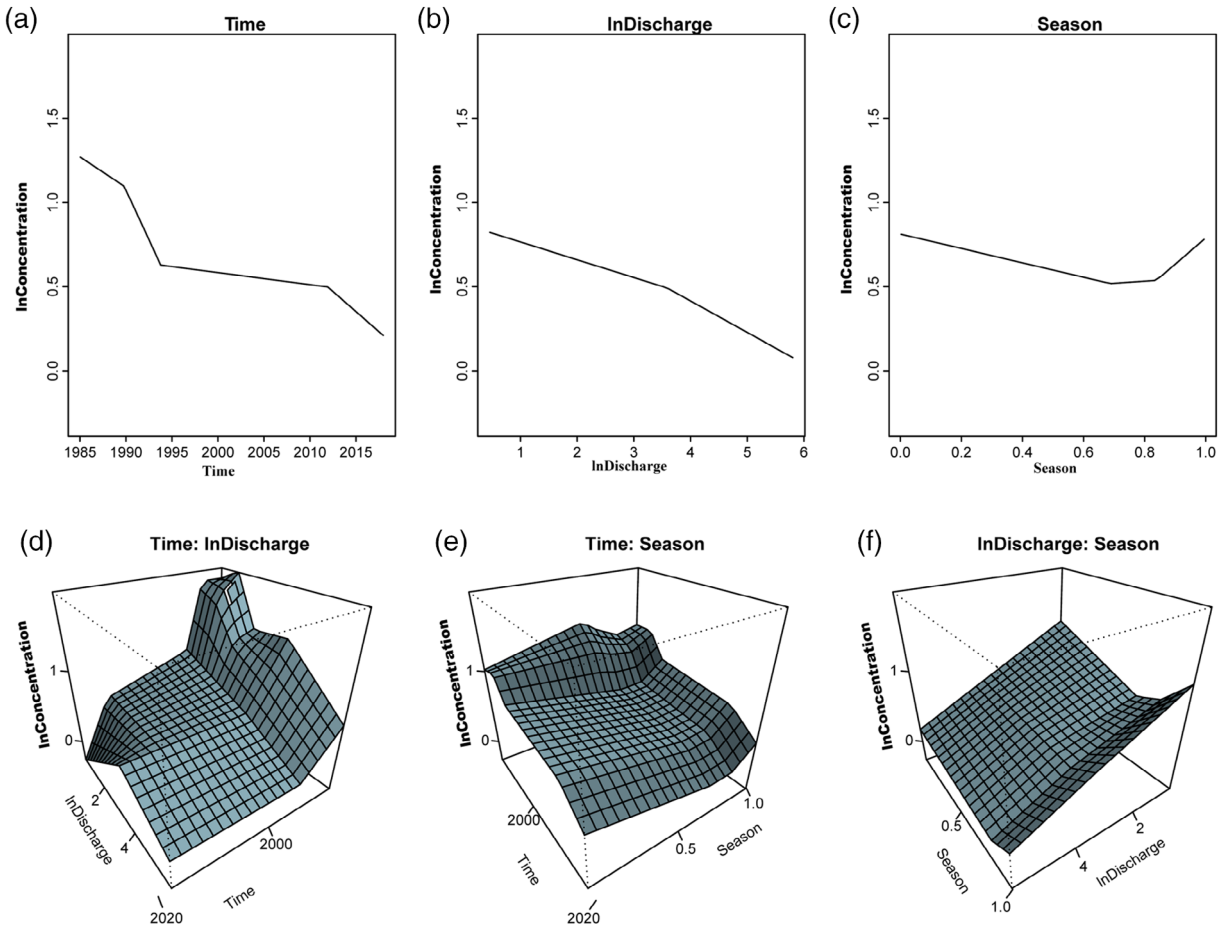

FIGURE 4 Main and interaction effects of MARS-EC model for TN in Patuxent River (1985–2017). (a) Concentration-time curve, (b) concentration-discharge curve, (c) concentration-season curve, (d) interaction between time and discharge, (e) interaction between time and season, and (f) interaction between discharge and season

changes between four time periods (1985–1989, 1990–1993, 1994–2011, and 2012–2017). Similarly, the lnConcentration–Season relation curves indicated differences between two time periods (1985–1993 and 1994–2017). The lnConcentration–lnDischarge curve showed lnConcentration progressively decreased with decreasing lnDischarge in three steps between 1985–1989, 1990–1993, and 1994–2011 before reversing this trend since 2012 (Figure 7). The lnConcentration–Season curve showed an inverted U-shape curve during the 1985–1993 period that has tended to reverse since 1994 (Figure 8).

River discharge is a major factor regulating constituent concentrations as it incorporates dilution and changing hydrologic flow paths (i.e., run-off vs. groundwater inputs) associated with storm events. In a watershed dominated by point source pollution, riverine constituent concentration might decrease with increasing discharge due to dilution (Chen et al., 2013). In a watershed dominated by nonpoint source pollution, riverine constituent concentration is generally assumed to be a power law function of discharge (Grizzetti et al., 2005; Huang, Zhang, & Lu, 2014). The shifts in the concentration versus discharge relationship in the Patuxent River appear to result from sewage treatment plant upgrades over the study period (Moyer et al., 2012). In accordance with the shifting trend pattern (Figure 6) and concentration-discharge and concentration-season curves with time (Figures 7 and 8), we posit that (a) TN concentration in the Patuxent River has decreased since 1985, (b) point-source pollution played a major role before 1995, and (c) nonpoint source pollution was dominant after 1995, particularly since 2010.

3.2 | Model flexibility

Constituent concentrations in a river segment are a function of discharge, time, and season, as well as several other environmental factors such as water temperature (TM), specific conductivity (SC), dissolved oxygen (DO), and pH (Chen et al., 2013; Huang, Zhang, & Lu, 2014). In order to improve model performance, it is necessary and meaningful to take into account additional explanatory variables. The exploratory variables in MARS-EC were set similar to ESTIMATOR and WRTDS. In 2004, the U.S. Geological Survey published a LOADEST program that incorporated parts of the original ESTIMATOR program code but had many enhancements, such as the ability to have user specified models with additional variables (e.g., turbidity

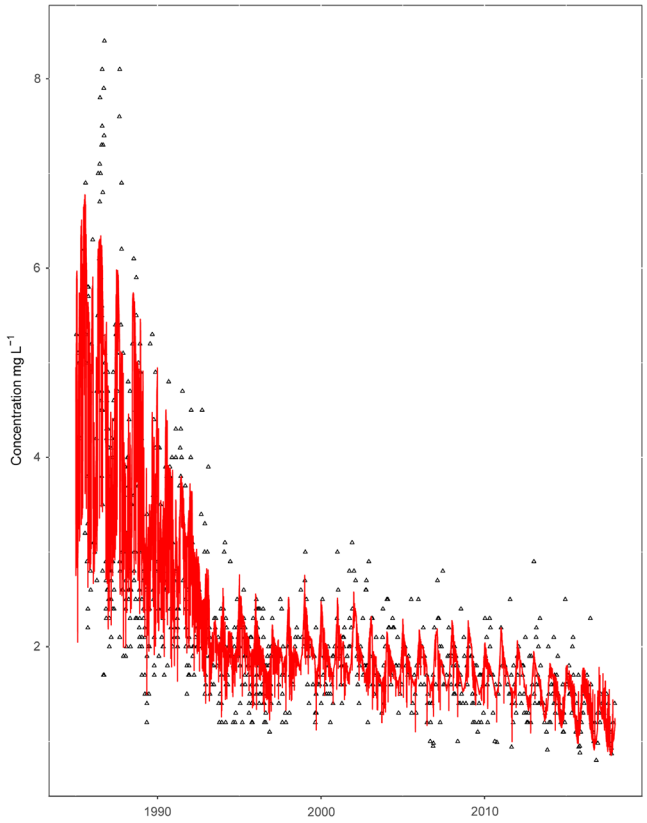

FIGURE 5 Measured (triangle) and daily estimated (line) TN concentrations in Patuxent River (1985–2017)

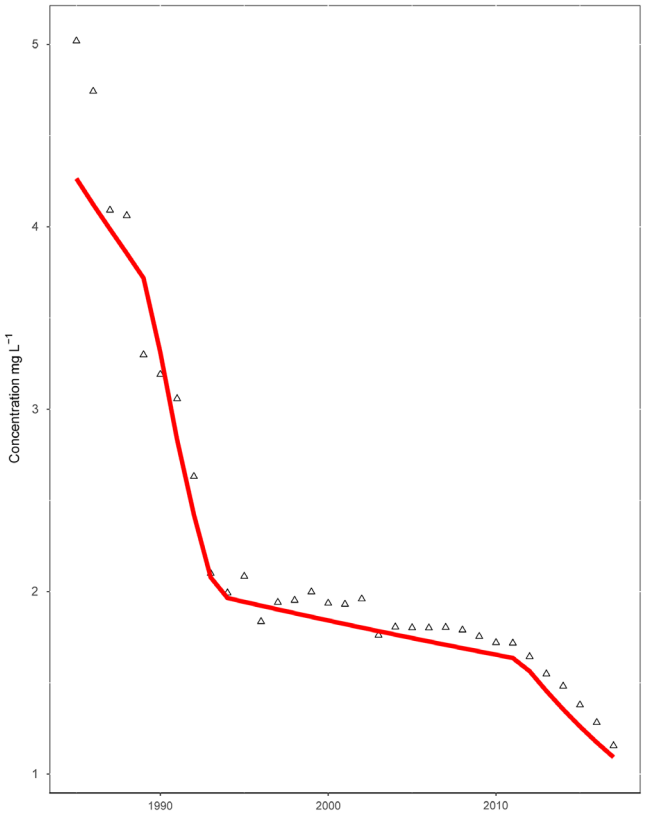

FIGURE 6 Yearly TN concentrations (triangle) and the trend patterns (line) in Patuxuent River (1985–2017)

and specific conductance; Runkel, Crawford, & Cohn, 2004). MARS is good at dealing with high dimensional data (Friedman & Roosen, 1995), and therefore, MARS-EC is very flexible in taking into account auxiliary exploratory variables. Herein, in addition to the standard predictor variables (lnTrend, lnDischarge, and Season) in Equation (5), we add three auxiliary predictor variables (TM, DO, and pH) to the MARS-EC model for estimating TN in the Patuxent River. Note that these auxiliary predictor variables are not daily measurements, but semi-monthly measurements taken in conjunction with TN concentrations at each monitoring date. The RSq increased from 0.80 to 0.82 with the addition of the auxiliary predictor variables, whereas the number of model terms decreased from 18 to 15 (Figures 2 and 9). Thus, the inclusion of auxiliary predictor variables provided a small improvement to model performance. Similarly, whether or not interaction effects are included in the MARS-EC model may also influence model performance. MARS-EC modelling in the "earth" packages can select variables automatically and thereby quickly determines the efficacy of including additional predictor variables and various interactions (Milborrow, 2018; Milborrow, 2019a). Given recent technological developments in monitoring, several environmental factors can be monitored in real time at a high frequency and at a greater number of monitoring stations. Inclusion of enhanced monitoring data could appreciably improve model performance. The flexibility of MARS-EC has the potential to improve the concentration fitting and estimation capabilities, as well as model interpretability, in many water resource applications (Figure 10). Therefore, the model flexibility of MARS-EC should be handled case by case with respect to the addition of auxiliary variables and inclusion of interaction effects. Interaction effects and auxiliary variables should only be included if they enhance model performance in a meaningful way.

3.3 | Model accuracy and uncertainty

Metrics for model accuracy and uncertainty are necessary to evaluate the efficacy of models for estimation of constituent concentrations for various water resource management activities. Model accuracy refers to the capability of the model to faithfully predict the true outcome (Casillas et al., 2003). The model accuracy of MARS-EC was evaluated by modelling nitrate plus nitrite concentrations for eight rivers in the Chesapeake Bay watershed during 1985–2017; model results were compared with those from the ESTIMATOR model. Note that the Potomac River was not used in this analysis because daily discharge records were not available for this site. Although we would ideally like to test the modelling efficacy of other important constituents, such as TN and total phosphorus (TP), data gaps for these parameters (>10 years) in about half of the Chesapeake Bay monitoring sites prevented their inclusion. The percentage of variance explained (in log space) and mean relative error (in real space) of the models were used to evaluate model accuracy. The percentage of

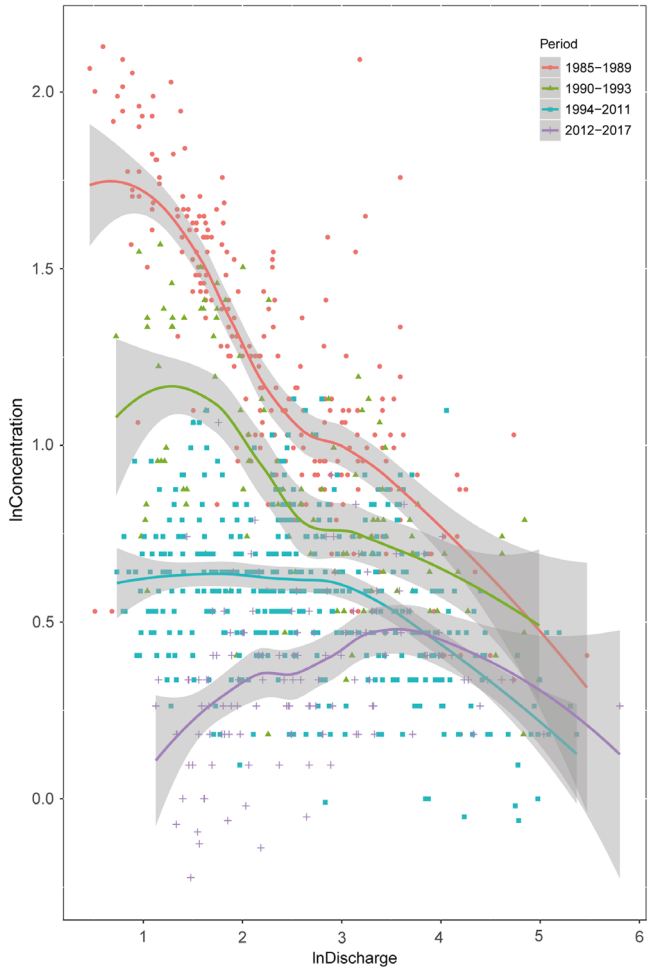

FIGURE 7 TN concentration-discharge curves for different time periods in the Patuxent River. The smooth curve is the loess smooth line and the shadow is the standard error

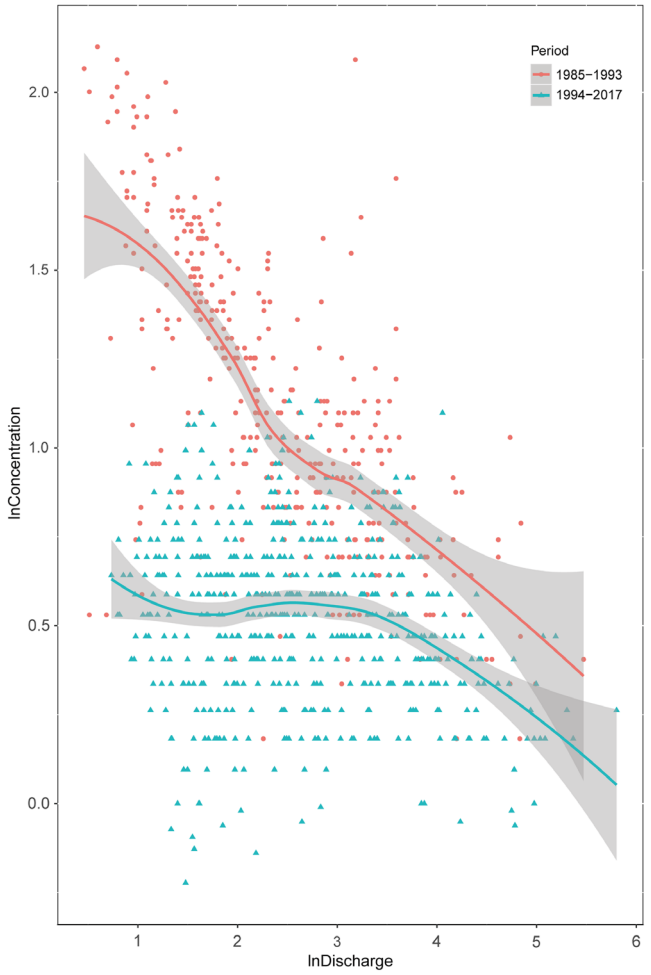

FIGURE 8 TN concentration-season curves for different time periods in the Patuxent River. The smooth curve is the loess smooth line, and the grey area is the standard error

variance explained value represents the closeness of estimated values to the measured values. The mean, minimum, and maximum percentage of variance explained values for MARS-EC models of the eight rivers were 57%, 38%, and 84%, compared with 45%, 15%, and 72% for ESTIMATOR models (Table 1). The mean relative error value represents the deviation of model estimates. The mean, minimum, and maximum values of mean relative error for MARS-EC models of the eight rivers were 38%, 17%, and 85%, compared with 50%, 22%, and 123% for ESTIMATOR models (Table 1). In general, accuracy of the MARS-EC models was superior to ESTIMATOR models. ESTIMATOR assumes the concentration-time and concentration-discharge curves to be either linear or quadratic (Cohn et al., 1989), and the shape of the curves is not allowed to change over time (Aulenbach et al., 2016). In contrast, MARS-EC uses multiple linear segmented regressions to approximate the underlying relationships between concentrations and explanatory variables, with interaction effects also being taken into account. Hence, the multiple linear segmented regression approach appears to enhance model accuracy of MARS-EC models compared with the ESTIMATOR approach for these datasets. It should be noted that RSq values of concentration estimating models are expected to be lower than flux estimating models because flux regression equations are inflated because discharge is an explanatory variable. Considering the complexities involving the transport and transformation of nutrients and in-stream assimilation (Chen et al., 2013), the model accuracy of the MARS-EC model is considered reasonable, and the model has good capability for accurately estimating riverine constituent concentrations.

Water quality modelling incorporates several uncertainties due to the complexity of hydro-biogeochemical interactions at the watershed scale (Defew, May, & Heal, 2013; Nguyen & Willems, 2016; Shrestha & Solomatine, 2008). We need to realize that both MARS-EC and ESTIMATOR had relatively poor performances on several rivers, such as the Susquehanna River and Choptank River, resulting in large uncertainties. Modelling uncertainty can be further classified into input, structural, parameter, and input components (Kasiviswanathan & Sudheer, 2017). Input uncertainty mainly arises from measurement and sampling uncertainties. Measurement errors are inherent in river discharge quantification and analytical errors in water quality analysis, while sampling uncertainty results from

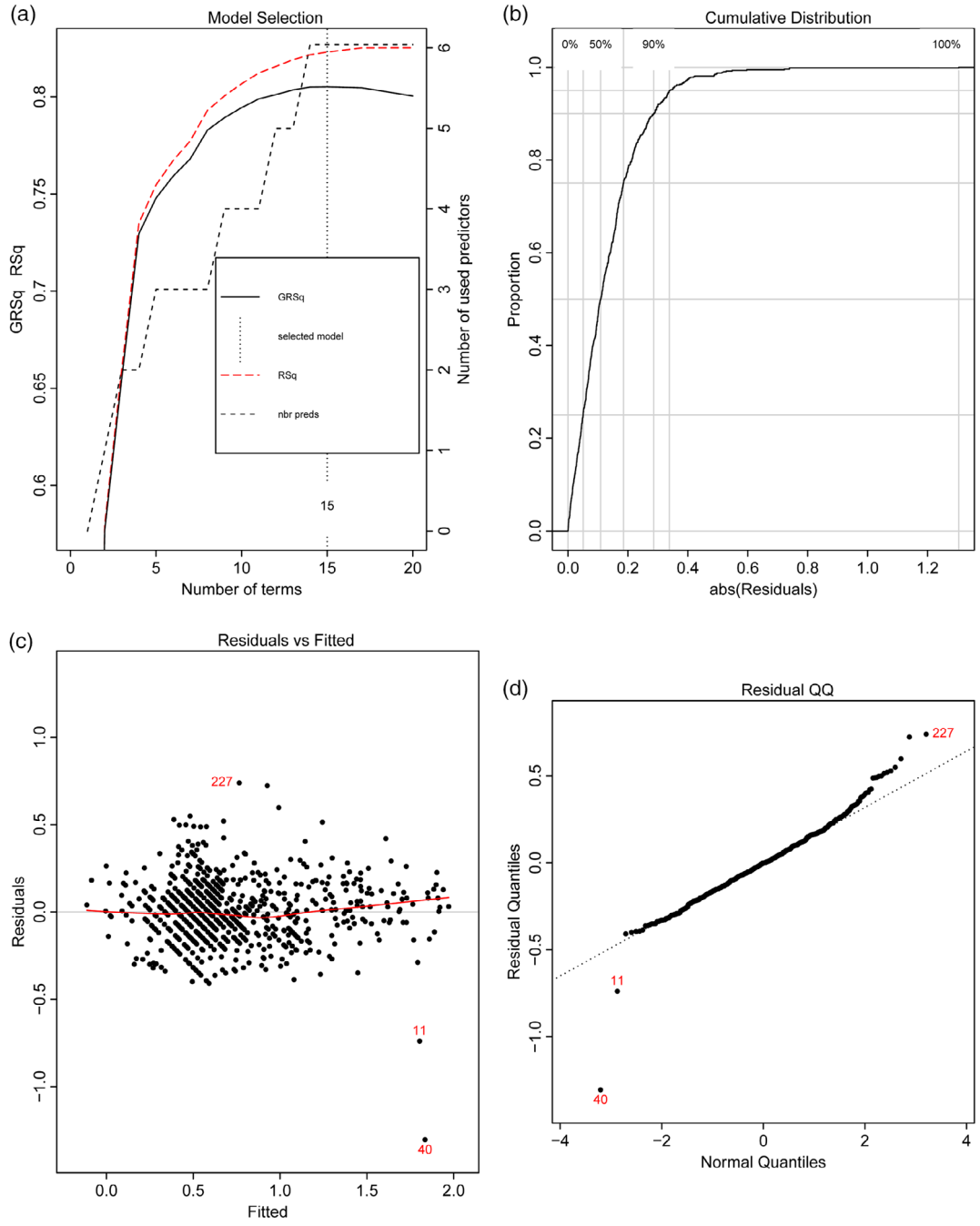

FIGURE 9 Model evaluation metrics of MARS-EC model with auxiliary predictor variables for TN in Patuxent River (1985–2017). (a) Model selection graph, (b) residuals versus fitted graph, (c) cumulative distribution graph, and (d) residual QQ graph

deficiencies in collecting a representative sample from a stream channel and across the full range of hydrologic conditions (Snelder, McDowall, & Fraser, 2017). For example, during the 1985–1995 periods, concentrations of nitrate plus nitrite in the Susquehanna River exhibited appreciable variability across monitoring dates (Figure 11). Relatively invariant datasets across different times, seasons, and discharge rates also hinder an effective model calibration. For sampling uncertainty, samples must be collected across the full range of hydrologic conditions (including targeted storm flow samples) to provide full representation of concentration-discharge conditions (Bowes, Smith, & Neal, 2009; Moyer et al., 2012). Model and parameter uncertainty originate from nonoptimum variable selection, model functions, residual variation, and model parameterization (Kasiviswanathan & Sudheer, 2017; Qin, Zhang, Zhong, & Yu, 2017). For structure and parameter uncertainty, both ESTIMATOR and MARS-EC share the premise that riverine constituent concentrations are functions of time, discharge, and season. A lnConcentration–lnDischarge relationship is a basic premise of regression-based models as illustrated by the nitrate plus nitrite versus discharge curve for the Susquehanna River during 1985–2017 (Figure 12). A linear regression

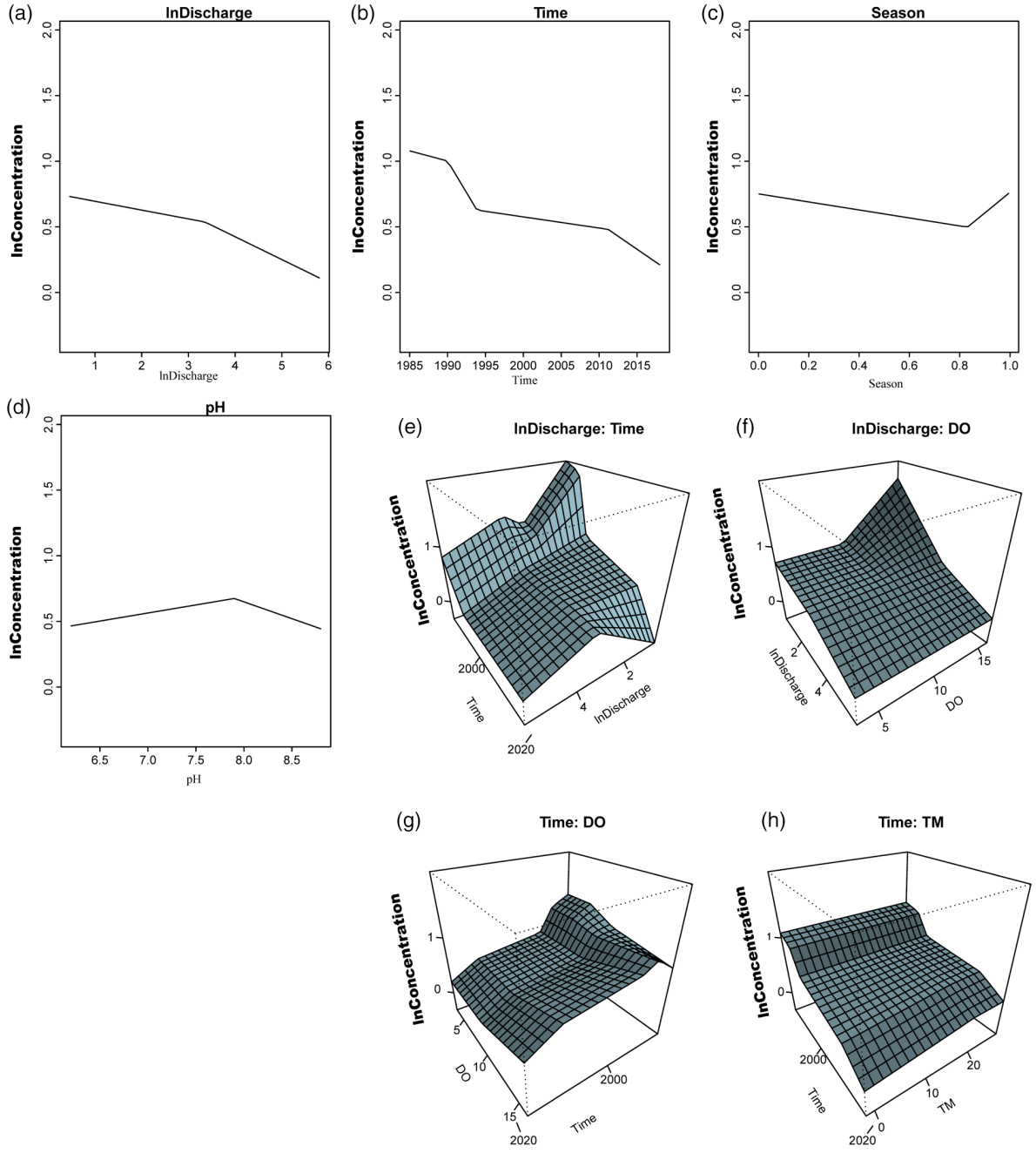

FIGURE 10 Main and interaction effects for MARS-EC model with auxiliary predictor variables for TN in the Patuxent River. (a) Concentration-discharge curve, (b) concentration-time curve, (c) concentration-season curve, (d) concentration-pH curve, (e) interaction between time and discharge, (f) interaction effect between discharge and DO, (g) interaction between time and DO, and (h) interaction between time and TM

based on lnDischarge explained 16% of the variance of lnConcentration, whereas a loess regression explained 32%. In this case, both ESTIMATOR and MARS-EC performed poorly due to the weak relationship between nitrate plus nitrite and discharge. Objectively speaking, no model will perform well in cases where the relationships between constituent concentrations and predictors are weak or irregular. Thus, uncertainty analysis is essential in modelling constituent concentrations to determine model efficacy for a given watershed of interest (Appling et al., 2016; Johnes, 2007). The more tools we have available, the better our chances of effectively modelling water quality dynamics for water resource protection and remediation.

3.4 | Summary

The major aim of this methodology-based research was to introduce a new MARS-EC approach that integrates the advantages of parametric and nonparametric models for estimating riverine constituent

River name	Percentage of variance explained (%)		Mean relative error (%)	
	MARS-EC	ESTIMOTER	MARS-EC	ESTIMOTER
Susquehanna	38	31	21	22
James	69	58	49	62
Rappahannock	74	61	85	123
Appomattox	41	64	47	66
Pamunkey	46	16	25	32
Mattaponi	47	15	36	50
Patuxent	84	72	17	22
Choptank	55	42	21	26
Mean	57	45	38	50

TABLE 1 Percentage of variance explained (in log space) and mean relative error (in real space) of MARS-EC and ESTIMATOR models for nitrate plus nitrite concentration estimates in eight rivers of Chesapeake Bay watershed

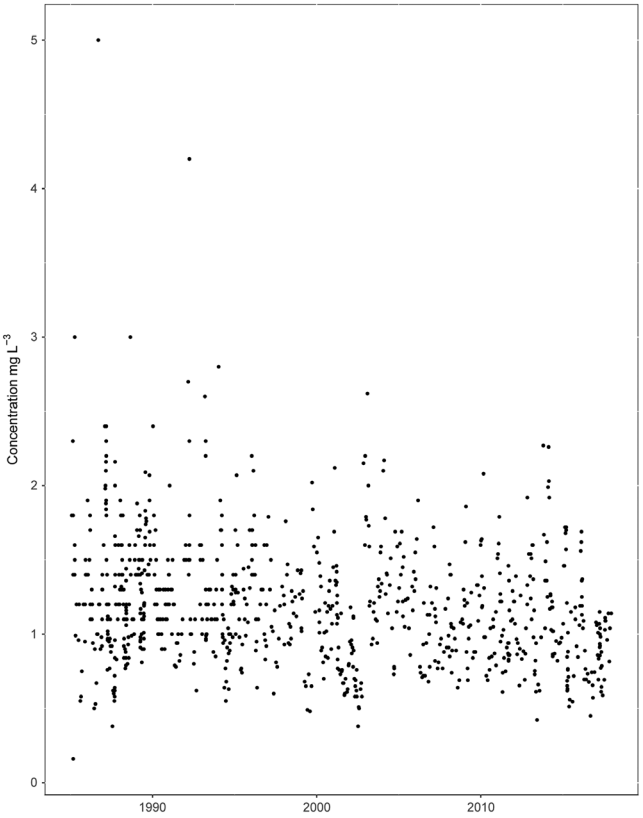

FIGURE 11 Concentrations of nitrate plus nitrite in Susquehanna River (1985–2017)

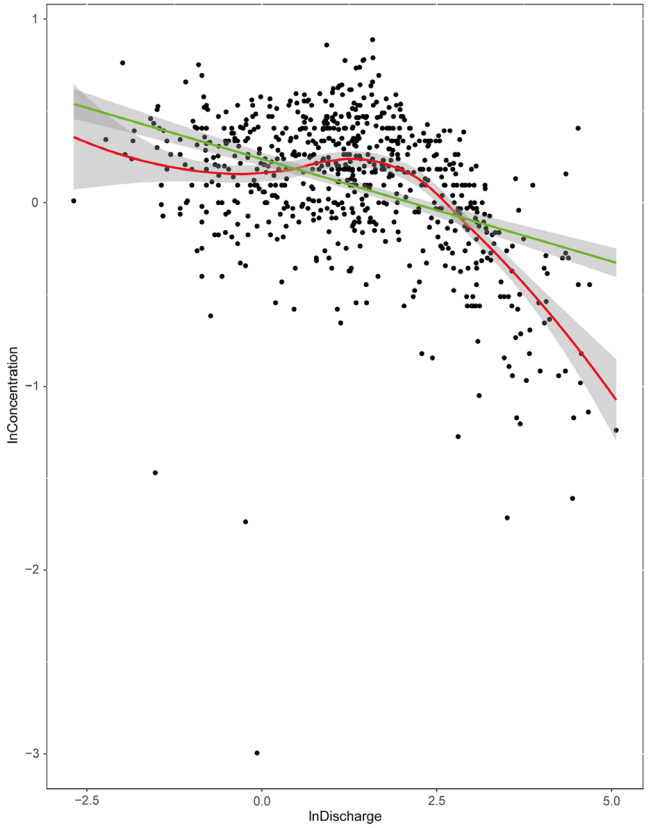

FIGURE 12 lnConcentration–lnDischarge curve of nitrate plus nitrite in Susquehanna River (1985–2017) determined by linear regression (green) loess smooth line (red)

concentrations. The various methods available for estimating riverine constituent concentrations have relative advantages and limitations. The MARS-EC approach developed in this study provides some obvious advantages. First, MARS-EC has a strong capacity to deal with high-dimensional analyses that allows for the investigation of several auxiliary explanatory variables to improve model performance, yet is flexible enough to optimize the "important" variables and interactions automatically. Second, MARS-EC does not assume an a priori concentration-predictor relationship but uses piecewise regression to approximate relationships and provides mathematical expressions and visual outputs to define break points (gradual or abrupt) in trend lines.

Third, MARS-EC takes into account interactions between predictor variables and has the ability to adjust concentration-predictor relationship curves, such as shifts in concentration-discharge curves, that may occur over long time periods. In general, MARS-EC is expected to have good potential for estimating constituent concentrations and trend analysis. Concentration-discharge and concentration-season relationships are important and perhaps sensitive indicators of biological and hydrological functioning in watersheds (Moatar, Abbott, Minaudo, Curie, & Pinay, 2017). Therefore, these relationships may change in response to land-use change and watershed management

practices over time (Moyer et al., 2012; Moatar et al., 2017; Zhang, 2018). Importantly, MARS-EC does not constrain concentration-discharge and concentration-season curves to be constant over long-term periods. This attribute is a big advantage of MARS-EC as detecting thresholds (i.e., change points) is a critical issue during this period of rapid environmental change.

We also identified several limitations of the newly developed MARS-EC approach that warrant further investigation. First, MARS-EC does not have the ability to directly deal with outlier/inaccurate data, an issue ubiquitous in water quality datasets (Oblinger, 1999). Second, MARS-EC is highly sensitive to data outliers and the lack of representative data across the entire range of environmental/hydrological conditions. Third, MARS-EC performance was weak, similar to other modelling approaches, when constituent concentrations showed little variability with respect to predictor variables. Furthermore, in contrast to mechanistic models, MARS-EC is a statistically based model that does not directly consider the influence of surface water and groundwater dynamics on riverine constituent concentrations. River discharge can be separated into surface water (direct run-off) and groundwater (baseflow) components, which often have very different chemical signatures. Although the concentration versus discharge curves incorporate some aspects of surface water-groundwater contributions, it may be beneficial to better quantify the contributions from surface water and groundwater on the integrated riverine concentrations. In any case, it is important to have several options to provide a simple and reliable estimate of water quality parameters to assist in watershed pollution control, management, and remediation (Huang, Zhang, & Lu, 2014; Santos, de Weys, Tait, & Eyre, 2013).

4 | CONCLUSIONS

The MARS-EC modelling approach achieves operational integration of several beneficial aspects of parametric and nonparametric regression-based methods and was demonstrated to be straightforward and effective for estimating riverine constituent concentrations. The modelling process of MARS-EC is flexible and allows consideration of several auxiliary explanatory variables, whereas variable selection and interactions are automatically optimized in the final model. MARS-EC is not constrained by the need for constant concentration-predictor curves but is rather able to identify shifts in these relationships (i.e., change points) due to environmental change using both mathematical expressions and visual outputs. The MARS-EC approach was developed to complement and supplement existing approaches for estimating riverine constituent concentrations. Each particular modelling approach may have advantages/disadvantages for a given watershed providing motivation for developing new approaches to enhance water quality management.

ACKNOWLEDGEMENTS

This work was supported by National Natural Science Foundation of China (41807495 and 41601554). We thank the editor and reviewers for their careful evaluation and exceptionally constructive comments.

CONFLICT OF INTERESTS

The authors declared that they have no conflicts of interest to this work. We declare that we do not have any commercial or associative interest that represents a conflict of interest in connection with the work submitted.

DATA AVAILABILITY STATEMENT

Supporting data for this study are openly available in USGS National Water Information System at http://waterdata.usgs.gov/nwis, reference number (USGS, 2019).

ORCID

Kun Mei https://orcid.org/0000-0002-4297-5637

REFERENCES

Appling, A. P., Leon, M. C., & McDowell, W. H. (2016). Reducing bias and quantifying uncertainty in watershed flux estimates: The R package loadflex. *Ecosphere*, *6*(12), 1–25.

Aulenbach, B. T., Burns, D. A., Shanley, J. B., Yanai, R. D., Bae, K., Wild, A. D., … Yi, D. (2016). Approaches to stream solute load estimation for solutes with varying dynamics from five diverse small watersheds. *Ecosphere*, *7*(6), 1–22.

Bowes, M. J., Smith, J. T., & Neal, C. (2009). The value of high resolution nutrient monitoring: A case study of the river Frome, Dorset, UK. *Journal of Hydrology*, *378*(1–2), 82–96.

Brakebill, J. W., Scott, W. A., & Schwarz, G. E. (2010). Sources of suspended-sediment flux in streams of the Chesapeake Bay watershed: A regional application of the SPARROW model. *Journal of the American Water Resources Association (JAWRA)*, *46*(4), 757–776.

Buckner, G. D., Choi, H., & Gibson, N. S. (2006). Estimating model uncertainty using confidence interval networks: Applications to robust control. *Journal of Dynamic Systems Measurement and Control*, *128*(3), 626–635.

Calamari, D., Nauen, C. E., & Naeve, H. (1987). Water quality problems in Africa: A brief report. *International Journal of Environmental Studies*, *29*(1), 3–8.

Casillas, J., Cordón, O., Herrera, F., & Magdalena, L. (2003). *Accuracy improvements to find the balance interpretability-accuracy in linguistic fuzzy modeling: an overview. Interpretability Issues in Fuzzy Modeling*. Berlin, Heidelberg: Springer-Verlag.

Chang, H. (2008). Spatial analysis of water quality trends in the Han River basin, South Korea. *Water Research*, *42*(13), 3285–3304.

Chen, D. J., Dahlgren, R. A., & Lu, J. (2013). A modified load apportionment model for identifying point and diffuse source nutrient inputs to rivers from stream monitoring data. *Journal of Hydrology*, *501*, 25–34.

Cohn, T. A., Caulder, D. L., Gilroy, E. J., Zynjuk, L. D., & Summers, R. M. (1992). The validity of a simple statistical model for estimating fluvial constituent loads: An empirical study involving nutrient loads entering Chesapeake Bay. *Water Resource Research*, *28*(9), 2353–2463.

Cohn, T. A., Delong, L. L., Gilroy, E. J., Gilroy, E. J., & Wells, D. K. (1989). Estimating constituent loads. *Water Resources Research*, *25*(25), 937–942.

DeCicco, L., Hirsch, R., Lorenz, D., et al., 2019. Retrieval Functions for USGS and EPA Hydrologic and Water Quality Data. https://cran.r-project.org/web/packages/dataRetrieval/dataRetrieval.pdf.

Defew, L. H., May, L., & Heal, K. V. (2013). Uncertainties in estimated phosphorus loads as a function of different sampling frequencies and common calculation methods. *Marine & Freshwater Research*, *64*(5), 373–386.

Deo, R. C., Kisi, O., & Singh, V. P. (2017). Drought forecasting in eastern Australia using multivariate adaptive regression spline, least square support vector machine and M5Tree model. *Atmospheric Research*, *184*, 149–175.

Dumont, E., Williams, R., Keller, V., VoÃ, A., & Tattari, S. (2012). Modelling indicators of water security, water pollution and aquatic biodiversity in Europe. *Hydrological Sciences Journal*, *57*(7), 1378–1403.

Fazel Zerandi, M. H., Zarinbal, M., Ghanbari, N., & Turksen, I. B. (2013). A new fuzzy functions model tuned by hybridizing imperialist competitive algorithm and simulated annealing. Application: Stock price prediction. *Information Science*, *222*, 213–228.

Friedman, J. H. (1991). Multivariate adaptive regression splines (with discussion). *Annals of Statistics*, *19*(1), 123–141. https://statistics.stanford.edu/research/

Friedman, J. H., & Roosen, C. B. (1995). An introduction to multivariate adaptive regression splines. *Statistical Methods in Medical Research*, *4*(3), 197–217.

Grizzetti, B., Bouraoui, F., de Marsily, G., & Bidoglio, B. (2005). A statistical method for source apportionment of riverine nitrogen loads. *Journal of Hydrology*, *304*(1–4), 302–315.

Hirsch, R. M. (2014). Large biases in regression-based constituent flux estimates: Causes and diagnostic tools. *Journal of the American Water Resources Association*, *50*(6), 1401–1424.

Hirsch, R. M., Archfield, S. A., & Cicco, L. A. D. (2015). A bootstrap method for estimating uncertainty of water quality trends. *Environmental Modelling & Software*, *73*, 148–166.

Hirsch, R.M., De Cicco, L.A., 2015. User guide to exploration and graphics for RivEr trends (EGRET) and dataRetrieval—R packages for hydrologic data. *US Geological Survey Techniques and Methods Book* 4, Chap. A10, 93. https://dx.doi.org/10.3133/tm4A10.

Hirsch, R. M., Moyer, D. L., & Archfield, S. A. (2010). Weighted regressions on time, discharge, and season (WRTDS), with an application to Chesapeake Bay River inputs. *Journal of the American Water Resources Association*, *46*(5), 857–880.

Huang, H., Chen, D. J., Zhang, B. F., Zeng, L., & Dahlgren, R. A. (2014). Modeling and forecasting riverine dissolved inorganic nitrogen export using anthropogenic nitrogen inputs, hydroclimate, and land-use change. *Journal of Hydrology*, *517*, 95–104.

Huang, H., Wang, Z., Xia, F., Shang, X., Liu, Y., Zhang, M., … Mei, K. (2017). Water quality trend and change-point analyses using integration of locally weighted polynomial regression and segmented regression. *Environmental Science & Pollution Research*, *24*(18), 15827–15837.

Huang, H., Zhang, B. F., & Lu, J. (2014). Quantitative identification of riverine nitrogen from point, direct runoff and base flow sources. *Water Science and Technology*, *70*(5), 865–870.

Johnes, P. (2007). Uncertainties in annual riverine phosphorus load estimation: Impact of load estimation methodology, sampling frequency, baseflow index and catchment population density. *Journal of Hydrology*, *332*(1–2), 241–258.

Kasiviswanathan, K. S., & Sudheer, K. P. (2017). Methods used for quantifying the prediction uncertainty of artificial neural network based hydrologic models. *Stochastic Environmental Research and Risk Assessment*, *31*(7), 1659–1670.

Khosravi, A., Mazloumi, E., Nahavandi, S., Creighton, D., & Lint, J. W. C. V. (2011). Prediction intervals to account for uncertainties in travel time prediction. *IEEE Transactions on Intelligent Transportation Systems*, *12*(2), 537–547.

Kisi, O. (2015). Pan evaporation modeling using least square support vector machine, multivariate adaptive regression splines and M5 model tree. *Journal of Hydrology*, *528*, 312–320.

Koba, K., & Bączek, T. (2013). The evaluation of multivariate adaptive regression splines for the prediction of antitumor activity of acridinone derivatives. *Medicinal Chemistry*, *9*(8), 1041–1050.

Leathwick, J. R., Elith, J., & Hastie, T. (2006). Comparative performance of generalized additive models and multivariate adaptive regression splines for statistical modelling of species distributions. *Ecological Modelling*, *199*(2), 188–196.

León, L. F., Soulis, E. D., Kouwen, N., & Farquhar, G. J. (2001). Nonpoint source pollution: A distributed water quality modeling approach. *Water Research*, *35*(4), 997–1007.

Lorca, P., & Juez, F. J. D. C. (2011). Bankruptcy forecasting: A hybrid approach using fuzzy c-means clustering and multivariate adaptive regression splines (MARS). *Expert Systems with Applications*, *38*(3), 1866–1875.

Milborrow, S., 2018. Build regression models using the techniques in Friedman's papers "Fast MARS" and "multivariate adaptive regression splines". http://www.milbo.users.sonic.net/earth/.

Milborrow S. 2019a. Notes on the earth package. http://www.milbo.org/doc/earth-notes.pdf.

Milborrow, S. 2019b. Plotting regression surfaces with plotmo. http://www.milbo.org/doc/plotmo-notes.pdf.

Moatar, F., Abbott, B. W., Minaudo, C., Curie, F., & Pinay, G. (2017). Elemental properties, hydrology, and biology interact to shape concentration-discharge curves for carbon, nutrients, sediment, and major ions. *Water Resources Research*, *53*, 1270–1287.

Moyer, D.L., Hirsch, R.M., Hyer, K.E., 2012. Comparison of two regression-based approaches for determining nutrient and sediment fluxes and trends in the Chesapeake Bay watershed: U.S. Geological Survey Scientific Investigations Report 2012–5244, 118 p, Available online at http://pubs.usgs.gov/sir/2012/5244/.

Nguyen, T. H., & Willems, P. (2016). The influence of model structure uncertainty on water quality assessment. *Water Resources Management*, *30*(9), 3043–3061.

Oblinger, C. J. (1999). New reporting procedures based on long-term method detection levels and some considerations for interpretations of water-quality data provided by the U.S. Geological Survey National Water Quality Laboratory. *Center for Integrated Data Analytics Wisconsin Science Center*, *33*(6), 693–697.

Oduro, S. D., Metia, S., Duc, H., Hong, G., & Ha, Q. P. (2015). Multivariate adaptive regression splines models for vehicular emission prediction. *Visualization in Engineering*, *3*, 13. https://viejournal.springeropen.com/articles/10.1186/s40327-015-0024-4

Ouyang, W., Cai, G., Huang, W., & Hao, F. (2016). Temporal-spatial loss of diffuse pesticide and potential risks for water quality in China. *Science of the Total Environment*, *541*(15), 551–558.

Qin, L. H., Zhang, M. Z., Zhong, S. H., & Yu, X. (2017). Model uncertainty in forest biomass estimation. *Acta Ecologica Sinica*, *37*(23), 7912–7919 (in Chinese).

Renwick, W. H., Vanni, J., Zhang, Q., & Patton, J. (2008). Water quality trends and changing agricultural practices in a Midwest U.S. watershed, 1994–2006. *Journal of Environmental Quality*, *37*(5), 1862–1874.

Runkel, R.L., Crawford, C.G., Cohn, T.A., 2004. *Load estimator (LOADEST): A FORTRAN program for estimating constituent loads in streams and rivers*. U.S. Geological Survey Techniques and Methods (Book 4, Chapter A5). Reston: U.S. Geological Survey, 75 p, https://pubs.usgs.gov/tm/2005/tm4A5/pdf/508final.pdf.

Saleh, A., & Du, B. (2004). Evaluation of SWAT and HSPF within BASINS program for the upper north Bosque River watershed in Central Texas. *American Society of Agricultural Engineers*, *47*(4), 1039–1049.

Santos, I. R., de Weys, J., Tait, D. R., & Eyre, B. D. (2013). The contribution of groundwater discharge to nutrient exports from a coastal catchment: Post-flood seepage increases estuarine N/P ratios. *Estuaries and Coasts*, *36*(1), 56–73.

Shipley, B., & Hunt, R. (1996). Regression smoothers for estimating parameters of growth analyses. *Annals of Botany*, *78*(5), 569–576.

Shrestha, D. L., & Solomatine, D. P. (2008). Data-driven approaches for estimating uncertainty in rainfall-runoff modelling. *International Journal of River Basin Management*, *6*(2), 109–122.

Snelder, T. H., McDowall, R. M., & Fraser, C. (2017). Estimation of catchment nutrient loads in New Zealand using monthly water quality

monitoring data. *Journal of the American Water Resources Association*, *51*(3), 158–178.

Yochum, S.E., 2000. A revised load estimation procedure for the Susquehanna, Potomac, Patuxent, and Choptank Rivers. U.S. Geological Survey Water-Resources Investigations Report 00-4156, p. 55. https://pubs.usgs.gov/wri/wri00-4156/wrir-00-4156.pdf.

Zhang, Q. (2018). Synthesis of nutrient and sediment export patterns in the Chesapeake Bay watershed: Complex and non-stationary concentration-discharge relationships. *Science of the Total Environment*, *618*, 1268–1283.

Zhang, Q., Brady, D. C., & Boynton, W. R. (2015). Long-term trends of nutrients and sediment from the nontidal Chesapeake watershed: An assessment of progress by river and season. *Journal of the American Water Resources Association*, *51*(6), 1534–1555.

Zhang, Q., Hirsch, R. M., & Ball, W. P. (2016). Long-term changes in sediment and nutrient delivery from Conowingo Dam to Chesapeake Bay: Effects of reservoir sedimentation. *Environmental Science & Technology*, *50*(4), 1877–1886.

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section at the end of this article.

How to cite this article: Huang H, Ji X, Xia F, et al. Multivariate adaptive regression splines for estimating riverine constituent concentrations. *Hydrological Processes*. 2019;1–15. https://doi.org/10.1002/hyp.13669